Applications of artificial intelligence to structure determination of organic compounds. XX*. Determination of groups attached to the skeleton of natural products using ^{13}C nuclear magnetic resonance spectroscopy

Gilberto do V. Rodrigues, Ivan P. de Arruda Campos and Vicente de P. Emerenciano
Instituto de Química, Universidade de São Paulo, Cx.P. 26077, 05599–970, São Paulo, SP, Brazil

Received 13 October 1995

Revised 2 May 1996

Abstract. A procedure for the identification of substituent groups (viz. angelate, tiglate, etc.) attached to any of the atoms in the conventional skeleton of a natural product is described. It consists in the use of the program MACRONO, which was developed for finding subspectra due to the carbons in the said substituent groups amid the raw ^{13}C NMR spectroscopic data from any given natural product (by means of comparisons of all possible subsets of the observed chemical shifts with those contained in an apposite database, built with literature ^{13}C NMR spectroscopic data regarding those groups). This procedure enables one to expunge the chemical shifts not due to skeletal carbons from the initial dataset, which then can be input to the expert system SISTEMAT, for skeletal identification.

1. Introduction

The structure of any natural product (NP) is conventionally divisible into three sub-units: (i) the skeletal atoms (SA); (ii) heteroatoms directly bound to the SA or unsaturations between them; and (iii) secondary carbon chains (SCC), usually bound to a SA through an ester or ether linkage.

The presence of SCCs in an NP often causes difficulties for some of the previously described modules of the expert system SISTEMAT [1–4] (those relating to ^{13}C NMR analysis), because the presence of signals due to non-skeletal carbon atoms may confuse the identification algorithms, leading to false results. Thus, we realized that it was of paramount importance for the correct operation of our expert system to have a module capable of pre-processing the raw data, in order to avoid this problem.

*For Part XIX, see [4].

In the present paper, the program MACRONO.EXE, a new module of SISTEMAT, is described. It operates by comparing the raw data obtained from an NP with information contained in an apposite database (built with ^{13}C NMR chemical shifts of the most common SCC) and finding the best matches, thus allowing one to trim the initial list of observed ^{13}C NMR chemical shifts.

In accordance with the classification system adopted in the conception of SISTEMAT, each atom belonging to the NP's structure is called NO (node). In this context, the SCCs present should be seen as a special kind of NO, akin to the super-atoms of DENDRAL SYSTEM [5], hereby called either macroNO. Already, the previous modules of SISTEMAT included a simplified type of database on macroNOs, stored in a separate file called MACSIS. However, the records within MACSIS, which store a three-letter alias (viz. TIG for tiglate), the number of each kind of atoms in the SCC and its carbons ^{13}C chemical shifts do not include enough information to be useful for our new module, the program MACRONO.EXE. On the other hand, as any major modification in MACSIS would imply in the recompilation of all other modules of SISTEMAT, we decided to provide the program MACRONO.EXE with its own tailor-made database on SCCs.

2. The database MACRONOS.DAT

For the purpose of testing the above program, an initial version of the new database (which is stored in a separate file named MACRONOS.DAT) was prepared with information pertaining to the 58 SCC most often found attached to the SA of sesquiterpene lactones [6]. The program, however, is of general use, in principle, provided that the MACRONOS.DAT be enlarged by the addition of information corresponding to more SCCs. This is a process in progress: as the next step we are now expanding the scope of that database to cover most of the SCCs known to occur connected to all types of terpenes. Each record of MACRONOS.DAT presents the following fields:

CODE: the ordinal representing of the position of the record within the database;
STATUS: Boolean value that informs whether the record is active or not; inactive records remain in the database, but are ignored until reactivated;
SHORT_NAME: the same three-letter alias used for this specific SCC in MACSIS;
NAME: a chemical name for the SCC (not necessarily the systematic name);
NDATA: the length (total number of characters) of the structure code vector;
NAT: total number of atoms in the SCC;
VECTOR: the structure code vector (coded in the same system used to represent the structure of the skeletons, but for option -8, which indicates a macroNO (and means that it cannot be connected to another SCC);
C13: a bidimensional matrix, where each line represents a carbon atom and each column stores two numbers, representing the lower- and upper-limiting values of the chemical shift for that carbon;
ND: the number of independent datasets of chemical-shifts represented by the C13 field (more datasets may be added, from the literature, as they become available).

3. The program MACRONO.EXE

Developed in PASCAL, MACRONO.EXE was designed both for the identification of those data due to any SCCs present in the NP molecule and for the maintenance of MACRONOS.DAT. The program

is menu-driven, having a user-interface similar to those used in the other modules of SISTEMAT. The menu entries are:

- Option (0) – to exit the program and return to DOS;
- Option (1) – to present a list of the files containing drawn macroNOs (as in DATASIS, the structure codification is carried out by drawing the SCC on the screen, and the drawing can be saved in a file for further use);
- Option (2) – to add new macroNOs (requests from the user all the information necessary for the building of a new record into MACRONOS.DAT and translates the drawn SCC into a structure code vector. If the SHORT_NAME provided already exists, it exhibits the stored data and prompts for the NMR chemical shifts, if different from those already present in the bank. The new NMR data are, then, included into the record and ND is incremented);
- Option (3) – to correct or modify records on the database. Allows changes in every field, except for VECTOR (If VECTOR is not correct, one should set STATUS = 0 and delete the record with Option (7).);
- Option (4) – to list all the data stored for any given macroNO in the database;
- Option (5) – to enter NMR data from the sample being investigated. Prompts for a file name. If the file does not already exist, one is created, and the user is asked to input the chemical shifts. The new file is then saved for further use;
- Option (6) – to search for possible macroNOs present in the sample being investigated (see below);
- Option (7) – to delete inactive records (STATUS = 0) from the MACRONOS.DAT; and
- Option (8) – to launch a secondary command processor without quitting the program.

The search for the macroNOs, selected with Option (6), is the fundamental operation of the program. Thus some further comments are due:

A representative macroNO record will usually include more than one independent ^{13}C NMR chemical shifts dataset from the literature (i.e., $\text{ND} > 1$). Hence, every carbon will be represented by more than one chemical shift value (which may be the same value, but as the chemical shifts are somewhat sensitive to the experimental conditions, they are more likely to be different, although similar values). To search for the macroNOs data within the sample's dataset, MACRONO.EXE defines a mean chemical shift ${}^m\delta(i)$, for each carbon C- i of the SCC. It then proceeds to compare every possible subset of the sample's dataset of chemical shifts with the list of ${}^m\delta(i)$ values for every SCC throughout the whole list of SCC stored in MACRONOS.DAT, while calculating, for every fit attempted, a value for the average error (ε) of the fit:

$$\varepsilon = \frac{\sum_{i=1}^n {}^m\delta(i) - \delta(i)_s}{n}.$$

When this process is ended, the program outputs a list of all the possible macroNOs (that is all macroNOs for which it found a best subset of the chemical shifts in the sample dataset that yielded $\varepsilon \leq \text{Gradient}$), together with the number of carbons in each macroNO and its associated ε value of fit. It must be pointed out that, before the search is begun, the user is prompted to input a numerical value for Gradient.

The number of carbons in the SCC is an important information because it helps the user in discarding those SCC which, although having yielded a low ε value, represent a proposal with too many carbon atoms (i.e., in a 20 signals sample dataset one may think of either a diterpene or a sesquiterpene bearing a SCC of five carbons, but surely macroNOs of more than six carbon atoms can be discarded, in principle).

4. Results and discussion

We have selected arbitrarily from the literature the ^{13}C NMR chemical shifts from three different sesquiterpene lactones substituted by SCCs of known structure. Those datasets were then provided as test inputs to MACRONO.EXE. The resulting outputs, together with the initial datasets and the corresponding structural formulas are presented below, as Tests 1–3 (see also Annex).

For Test 1, seven possible macroNOs were proposed by the program.

In view of the fact that the initial dataset presents 23 signals and that the sample is a terpenoid, the possibilities can be narrowed to three hypotheses: (i) a diterpene bearing a SCC with three carbons (discarded for no such macroNO appears in the list); (ii) a nor-sesquiterpene presenting two different SCCs, one having five and the other four carbons; (iii) a sesquiterpene substituted by a SCC of eight carbons (see Table 1).

The decision in favour of the hypothesis (iii), however, is only possible after inspection of the ^1H NMR spectrum from the sample.

The analysis of the output for Tests 2 and 3 is simpler, because it is possible to assign correctly the SCCs present in those samples on the basis of the average errors and carbon numbers only.

It must be pointed out that the major contributor to the observed average errors stems from the differences in the chemical shifts of the carbonyl carbons (or of the ether carbon in the case of the glycosylated lactone), because of our initial decision to build the records in MACRONOS.DAT with ^{13}C NMR data for the free acids and sugars. These results could be better still if we had instead used the data from (say) the ethyl derivatives of the SCCs (the signals for the ethyl group being, of course, excluded). Even so, as we had already collected the datasets for the free macroNOs, it would add an unnecessary extra time to the test of the underlying ideas described in this paper, not justifiable when it is apparent that the accuracy of the results is already satisfactory. However, in the building of the next, more comprehensive version of MACRONOS.DAT, the use of data from ethyl derivatives shall be a standard procedure.

In conclusion, it should be mentioned that, after the identification and subsequent removal of the data due to the SCCs present in the molecules, the other modules of SISTEMAT did, in fact, achieve quicker and more reliable NP skeleton identifications for all the test compounds (data not shown).

Acknowledgment

Thanks are due to CNPq and FAPESP for financial support.

Annex

Test 1 [7] (Fig. 1)

^{13}C NMR: 170,7(1); 170,4(1); 140,5(1); 139,4(1); 127,2(1); 121,4(3); 101,0(2); 76,6(2); 76,5(2); 75,9(2); 74,0(2); 70,3(2); 68,4(3); 63,6(3); 42,5(3); 41,3(2); 37,1(3); 34,0(1); 28,4(3); 27,7(3); 27,0(4); 20,9(4); 18,5(3).

Number of Signals: 23

Fig. 1.

Results for Test 1

Code	macroNO	average error	no. of carbons
0	acetate	0.750	2
8	1,2-diOH-et-acrylate	1.600	6
24	2-OH-et-acrylate	1.6000	5
29	4-OH-metacrylate	1.900	4
43	metacrylate	2.400	4
56	6-ac-glucose	1.925	8
57	glucose	1.708	6

Gradient: 5.000 ppm.

Fig. 2.

Results for Test 2

Code	macroNO	average error	no. of carbons
0	acetate	0.450	2
48	propionate	2.800	3

Gradient: 5.000 ppm.

Test 2 [8] (Fig. 2)

^{13}C NMR: 173,6(1); 169,8(1); 163,7(1); 149,4(1); 137,5(1); 127,1(1); 123,7(2); 114,9(3); 79,5(2); 77,2(2); 75,2(2); 54,7(3); 37,0(3); 26,2(3); 24,2(3); 21,1(4); 11,4(4)

Number of Signals: 17

Test 3 [9] (Fig. 3)

^{13}C NMR: 170,1(1); 169,5(1); 166,0(1); 141,6(2); 138,0(1); 131,9(1); 127,9(1); 120,8(2); 118,3(3); 80,6(2); 77,7(2); 76,2(2); 59,6(3); 47,1(2); 46,4(2); 42,6(1); 29,4(3); 23,0(3); 21,3(4); 20,9(4); 12,6(4); 8,7(4)

Number of Signals: 22

Fig. 3.

Results for Test 3

Code	macroNO	average error	no. of carbons
0	acetate	0.300	2
3	4-ac-angelate	2.243	7
20	4-OH-angelate	1.340	5
35	4-OH-tiglate	0.880	5
44	2-me-butyrate	3.020	5
55	tiglate	1.280	5

Gradient: 5.000 ppm.

Table 1

The most frequent macroNOs attached to sesquiterpene lactones

Structure	Name	Structure	Name
	acetate		2-ac-et-acrylate
	6-ac-glucose		3-ac-isovalerate
	4-ac-angelate		5-ac-tiglate

Table 1
(Continued)

Structure	Name	Structure	Name
ac-sarracinate		2,3-diac-2-me-butyrate
angelate		1,2-diac-et-acrylate
2,3-diOH-isobutyrate		2,3-diOH-2-me-butyrate
4,5-diOH-tiglate		ep-angelate
2,3-ep-n-butyrate		ep-me-acrylate
ep-sarracinate		furoate
glucose		hidrocinnamate

Table 1
(Continued)

Structure	Name	Structure	Name
2-OH-3-ac-2-me-butyrate		4-OH-5-ac-tiglate
4-OH-angelate		2-OH-3-Cl-isobutyrate
5-OH-3-SH-sarracinate		3-OH-2-eto-isobutyrate
2-OH-et-acrylate		4-OH-5[5-OH-tig]-tiglate
2-OH-isobutyrate		4-OH-isobutyrate
3-OH-isovalerate		4-OH-metacrylate
3-OH-2-me-butyrate		2-OH-3-oxo-2-me-butyrate

Table 1
(Continued)

Structure	Name	Structure	Name
	P-OH-phenyl-acetate		4-OH-sarracinate
	4[3-OH-ste]-tiglate		4-OH-tiglate
	5-OH-tiglate		5-tig-tiglate
	5[5-OH-tig]-tiglate		isobutyrate
	isovalerate		
	linoleate		
	lonolenate		methacrylate
	2-me-butyrate		3-me-2,3-diOH-valerate
	palmitate		piperidine

Table 1
(Continued)

Structure	Name	Structure	Name
propionate		cis-3-Sme-acrylate
trans-3-Sme-acrylate		sarracinate
senecionate		stearate
5-ste-isovalerate		tiglate

References

- [1] J.P. Gastmans, M. Furlan, M.N. Lopes, J.H.G. Borges and V.P. Emerenciano, *Quim. Nova* **13** (1990), 10.
- [2] J.P. Gastmans, M. Furlan, M.N. Lopes, J.H.G. Borges and V.P. Emerenciano, *Quim. Nova* **13** (1990), 75.
- [3] V.P. Emerenciano, G.V. Rodrigues, P.A.T. Macari, S.A. Vestri, J.H.G. Borges, J.P. Gastmans and D.L.G. Fromanteau, *Spectroscopy* **12** (1994), 91.
- [4] V.P. Emerenciano, L.D. Melo, G.V. Rodrigues and J.P. Gastmans, *Spectroscopy* **13** (1997), 181.
- [5] R.E. Carhart, D.H. Smith, H. Brown and C. Djerassi, *J. Am. Chem. Soc.* **97** (1975), 5755.
- [6] F.C. Seaman, *Botan. Rev.* **48** (1982), 121.
- [7] A.L. Pérez, L. Nava and A. R. Vivar, *Phytochemistry* **26** (1987), 765.
- [8] S.M. Khafagy, A.A. Seif El-Din, J. Jakupovic, C. Zedero and F. Bohlmann, *Phytochemistry* **27** (1988), 1125.
- [9] J. Jakupovic, M. Grenz, F. Bohlmann, D. Wasshausen and R.M. King, *Phytochemistry* **28** (1989), 1937.